Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

Variance inflation factors in the analysis of complex survey data

by Dan Liao and Richard Valliant

June 2012

Statistics Canada Statistique Canada

Canada

Survey Methodology, June 2012
Vol. 38, No. 1, pp. 53-62
Statistics Canada, Catalogue No. 12-001-X

Variance inflation factors in the analysis of complex survey data

Dan Liao and Richard Valliant [1]

Abstract

Survey data are often used to fit linear regression models. The values of covariates used in modeling are not controlled as they might be in an experiment. Thus, collinearity among the covariates is an inevitable problem in the analysis of survey data. Although many books and articles have described the collinearity problem and proposed strategies to understand, assess and handle its presence, the survey literature has not provided appropriate diagnostic tools to evaluate its impact on regression estimation when the survey complexities are considered. We have developed variance inflation factors (VIFs) that measure the amount that variances of parameter estimators are increased due to having non-orthogonal predictors. The VIFs are appropriate for survey-weighted regression estimators and account for complex design features, *e.g.*, weights, clusters, and strata. Illustrations of these methods are given using a probability sample from a household survey of health and nutrition.

Key Words: Cluster sample; Collinearity diagnostics; Linearization variance estimator; Survey-weighted least squares; Stratified sample.

1. Introduction

Collinearity of predictor variables in a linear regression refers to a situation where explanatory variables are correlated with each other. The terms, multicollinearity and ill conditioning, are also used to denote the same situation. Collinearity is worrisome for both numerical and statistical reasons. The estimates of slope coefficients can be numerically unstable in some data sets in the sense that small changes in the $\mathbf{X}$'s or the $\mathbf{Y}$'s can produce large changes in the values of these estimates. Statistically, correlation among the predictors can lead to slope estimates with large variances. In addition, when $\mathbf{X}$'s are strongly correlated, the R^2 in a regression can be large while the individual slope estimates are not statistically significant. Even if slope estimates are significant, they may have signs that are the opposite of what are expected (Neter, Kutner, Wasserman and Nachtsheim 1996). Collinearity may also affect forecasts (Smith 1974; Belsley 1984).

In experimental designs, it may be possible to create situations where the explanatory variables are orthogonal to each other. But, in many surveys, variables that are substantially correlated are collected for analysis. For example, total income and its components (*e.g.*, wages and salaries, capital gains, interest and dividends) are collected in the Panel Survey of Income Dynamics (http://psidonline.isr.umich.edu/) to track economic well-being over time. When one explanatory variable is a linear combination of the others, this is known as perfect collinearity (or multicollinearity) and is easy to identify. Cases that are of interest in practice are ones where collinearity is less than perfect but still affects the precision of estimates (Kmenta 1986, section 10.3).

Although there is a substantial literature on regression diagnostics for non-survey data, there is considerably less for survey data. A few articles in the last decade introduced techniques for the evaluation of the quality of regression on complex survey data, mainly on identifying influential points and influential groups with abnormal data values or survey weights. Elliot (2007), for instance, developed Bayesian methods for weight trimming of linear and generalized linear regression estimators in unequal probability-of-inclusion designs. Li (2007a, b); Li and Valliant (2009, 2011) adapted and extended a series of traditional diagnostic techniques to regression on complex survey data, mainly on identifying influential observations and influential groups of observations. Li's research covers residuals and leverages, DFBETA, DFBETAS, DFFIT, DFFITs, Cook's Distance and the forward search approach. Although an extensive literature in applied statistics provides valuable suggestions and guidelines for data analysts to diagnose the presence of collinearity (*e.g.*, Farrar and Glauber 1967; Theil 1971; Belsley, Kuh and Welsch 1980; Fox 1984; Belsley 1991), none of this research touches upon diagnostics for collinearity when fitting models with survey data.

The variance inflation factor (VIF) described in section 2, is one of the most popular conventional collinearity diagnostic techniques, and is mainly aimed at ordinary or weighted least squares regressions. A VIF measures the inflation of the variance of a slope estimate caused by the nonorthogonality of the predictors over and above what the variance would be with orthogonality. In section 3, we consider the case of an analyst who estimates model parameters using survey-weighted least squares (SWLS) and derive VIFs appropriate to SWLS. The components of the VIF can be estimated using the ingredients of a variance

1. Dan Liao, RTI International, 701 13th Street, N.W., Suite 750, Washington DC, 20005. E-mail: dliao@rti.org; Richard Valliant, University of Michigan and University of Maryland, Joint Program in Survey Methodology, 1218 Lefrak Hall, College Park, MD, 20742.

estimator that is in common usage in software packages for analyzing survey data. In the case of linear regression, a type of sandwich variance estimator will estimate both the model variance and design variance of the SWLS slope estimator. As we will show in section 3, the model or design variance of $\hat{\beta}_k$, an estimator of slope associated with the predictor $\mathbf{x}_k$, is inflated somewhat when different predictors are correlated with each other compared to what the variance would be if $\mathbf{x}_k$ were orthogonal to the other predictors. The measure of inflation, the VIF, is composed of terms that must be estimated from the sample. Our approach has been to substitute estimators that have both a model and design interpretation as described in section 3.5.

The fourth section presents an empirical study using data from the United States National Health and Nutrition Examination Survey. The application of our new approach is demonstrated and the newly-derived VIF values for SWLS are compared to the ones for OLS or WLS, which can be obtained from the standard statistical packages. The comparisons show that VIF values are different for different regression methods and a VIF specific to complex sample should be used to evaluate the harmfulness of collinearity in the analysis of survey data.

2. Collinearity diagnostics in ordinary least squares estimation

Suppose the sample s has n units, on each of which p $\mathbf{x}$'s or predictors and one analysis variable Y are observed. The standard linear model in a nonsurvey setting is $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{Y}$ is an $n \times 1$ vector of observations on a response or dependent variable; $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_p)$ is an $n \times p$ design matrix of fixed constants with $\mathbf{x}_k$, the $n \times 1$ vector of values of explanatory variable k for the n sample units; $\boldsymbol{\beta}$ is a $p \times 1$ vector of parameters to be estimated; and $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of statistically independent error terms with zero mean and constant variance σ^2. We assume, for simplicity, that $\mathbf{X}$ has full column rank. The ordinary least squares (OLS) estimate of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$, for which the model variance is $\mathrm{Var}_M(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$. Here, we use the subscript M to denote expectation under the model.

Collinearities of explanatory variables inflate the model variance of the regression coefficients compared to having orthogonal $\mathbf{X}$'s. This effect can be seen in the formula for the variance of a specific estimated non-intercept coefficient $\hat{\beta}_k$ (Theil 1971),

$$\mathrm{Var}_M(\hat{\beta}_k) = \frac{\sigma^2}{\sum_{i \in s} x_{ik}^2} \frac{1}{1 - R_k^2} \qquad (1)$$

where R_k^2 is the square of the multiple correlation from the regression of the k^{th} column of $\mathbf{X}$ on the other columns. This R-square defined as $R_k^2 = \hat{\beta}_{(k)}\mathbf{X}_{(k)}^T\mathbf{X}_{(k)}\hat{\beta}_{(k)} / \mathbf{x}_k^T\mathbf{x}_k$, where $\hat{\beta}_{(k)}$ is OLS estimate of the slope when $\mathbf{x}_k$ is regressed on the other $\mathbf{x}$'s and $\mathbf{X}_{(k)}$ is the $\mathbf{X}$ matrix with the k^{th} column removed. The term $\sigma^2 / \sum x_{ik}^2$ is the model variance of $\hat{\beta}_k$ if the k^{th} predictor were orthogonal to all the other predictors. The value of R_k^2 may be nonzero because the k^{th} predictor is correlated with one other explanatory variable or because of a more complex pattern of dependence between $\mathbf{x}_k$ and several other predictors. Consequently, the collinearity between $\mathbf{x}_k$ and some other explanatory variables can result in the inflation of the variance of $\hat{\beta}_k$ beyond what would be obtained with orthogonal $\mathbf{X}$'s. The second term in (1), $(1 - R_k^2)^{-1}$, is called the variance-inflation factor (VIF) (Theil 1971).

A basic reference on collinearity and other OLS diagnostics is Belsley *et al.* (1980). Collinearity diagnostics are covered in many other textbooks including Fox (1984) and Neter *et al.* (1996). In some cases, it is desirable to weight cases differentially in a regression analysis to incorporate a nonconstant residual variance. This form of weighting is model-based and is called weighted least squares (WLS). Most of current statistical software packages, (*e.g.*, SAS, Stata, S-Plus and R), use $(1 - R_{k(\mathrm{WLS})}^2)^{-1}$ as VIF for WLS, where $R_{k(\mathrm{WLS})}^2$ is the square of the multiple correlation from the WLS regression of the k^{th} column of $\mathbf{X}$ on the other columns. Fox and Monette (1992) also generalized this concept of variance inflation as a measure of collinearity to a subset of parameters in $\mathbf{b}$ and derived a *generalized variance-inflation factor* (GVIF). Furthermore, some interesting work has developed VIF-like measures, such as *collinearity indices* in Steward (1987) that are simply the square roots of the VIFs and *tolerance* defined as the inverse of VIF in Simon and Lesage (1988).

3. VIF in survey weighted least squares regression

3.1 Survey-weighted least squares estimators

Suppose the underlying structural model in the superpopulation is $\mathbf{Y} = \mathbf{X}^T\boldsymbol{\beta} + \mathbf{e}$, where the error terms in the model have a general variance structure $\mathbf{e} \sim (0, \sigma^2\mathbf{V})$ with known $\mathbf{V}$ and σ^2. Define $\mathbf{W}$ to be the diagonal matrix of survey weights. We assume throughout that the survey weights are constructed in such a way that they can be used for estimating finite population totals. The survey weighted least squares (SWLS) estimator is $\hat{\boldsymbol{\beta}}_{\mathrm{SW}} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y}$, assuming $\mathbf{X}^T\mathbf{W}^{-1}\mathbf{X}$ is invertible. Fuller (2002) describes the properties of this estimator.

The estimate $\hat{\boldsymbol{\beta}}_{\mathrm{SW}}$ is a model unbiased estimator of $\boldsymbol{\beta}$ under the model $\mathbf{Y} = \mathbf{X}^T\boldsymbol{\beta} + \mathbf{e}$ regardless of whether

$\text{Var}_M(\mathbf{e}) = \sigma^2\mathbf{V}$ is specified correctly or not, and is approximately design-unbiased for the census parameter $\mathbf{B}_U = (\mathbf{X}_U^T\mathbf{X}_U)^{-1}\mathbf{X}_U^T\mathbf{Y}_U$, in the finite population of N units. The subscript U stands for the finite population, $\mathbf{Y}_U = (Y_1, ..., Y_N)^T$, and $\mathbf{X}_U = (\mathbf{x}_1, ..., \mathbf{x}_p)$ with $\mathbf{x}_k$ as the $N \times 1$ vector of values for covariate k.

3.2 Model variance of coefficient estimates

The model variance of the parameter estimator $\hat{\boldsymbol{\beta}}_{\text{SW}}$, assuming $\text{Var}_M(\mathbf{e}) = \sigma^2\mathbf{V}$, can be expressed as

$$\begin{aligned}\text{Var}_M(\hat{\boldsymbol{\beta}}_{\text{SW}}) &= \sigma^2(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\tilde{\mathbf{V}}\tilde{\mathbf{X}}(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1} \\ &= \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}\sigma^2 = \mathbf{G}\sigma^2, \end{aligned} \quad (2)$$

where $\tilde{\mathbf{X}} = \mathbf{W}^{1/2}\mathbf{X}$, $\tilde{\mathbf{V}} = \mathbf{W}^{1/2}\mathbf{V}\mathbf{W}^{1/2}$, $\mathbf{A} = \tilde{\mathbf{X}}^T\tilde{\mathbf{X}}$, $\mathbf{B} = \tilde{\mathbf{X}}^T\tilde{\mathbf{V}}\tilde{\mathbf{X}}$, and $\mathbf{G} = \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$.

If the columns of $\mathbf{X}$ are orthogonal, then $\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} = \text{diag}(\tilde{\mathbf{x}}_k^T\tilde{\mathbf{x}}_k)$ and $\mathbf{A}^{-1} = \text{diag}(1/\tilde{\mathbf{x}}_k^T\tilde{\mathbf{x}}_k)$, where $\tilde{\mathbf{x}}_k = \mathbf{w}_k^{1/2}\mathbf{x}_k$. The ij^{th} element of $\mathbf{G}$ then becomes $\tilde{\mathbf{x}}_i^T\tilde{\mathbf{V}}\tilde{\mathbf{x}}_j/(\tilde{\mathbf{x}}_i^T\tilde{\mathbf{x}}_i)^2$. Thus, when the $\mathbf{X}$'s are orthogonal, the model variance of $\hat{\beta}_{\text{SW}_k}$ is

$$\text{Var}_M(\hat{\beta}_{\text{SW}_k}) = \sigma^2\tilde{\mathbf{x}}_k^T\tilde{\mathbf{V}}\tilde{\mathbf{x}}_k/(\tilde{\mathbf{x}}_k^T\tilde{\mathbf{x}}_k)^2, \quad (3)$$

a fact we will use later. More generally, the model variance of $\hat{\beta}_{\text{SW}_k}$, the coefficient estimate for the k^{th} explanatory variable, is

$$\text{Var}_M(\hat{\beta}_{\text{SW}_k}) = \mathbf{i}_k'\,\text{Var}_M(\hat{\boldsymbol{\beta}}_{\text{SW}})\mathbf{i}_k = \sigma^2\mathbf{i}_k'\mathbf{G}\mathbf{i}_k = \sigma^2 g^{kk} \quad (4)$$

where $\mathbf{i}_k$ is a $p \times 1$ vector with 1 in position k and 0's elsewhere, and g^{kk} is the k^{th} diagonal element of matrix $\mathbf{G}$.

3.3 Model-based VIF

As shown in Appendix A, the model variance of $\hat{\beta}_{\text{SW}_k}$ in (4) can be written as:

$$\text{Var}_M(\hat{\beta}_{\text{SW}_k}) = g^{kk}\sigma^2 = \frac{\zeta_k\rho_k}{1 - R^2_{\text{SW}(k)}}\frac{\sigma^2\tilde{\mathbf{x}}_k^T\tilde{\mathbf{V}}\tilde{\mathbf{x}}_k}{(\tilde{\mathbf{x}}_k^T\tilde{\mathbf{x}}_k)^2}, \quad (5)$$

where

$$\zeta_k = \frac{\tilde{\mathbf{e}}_{xk}^T\tilde{\mathbf{V}}\tilde{\mathbf{e}}_{xk}}{\tilde{\mathbf{e}}_{xk}^T\tilde{\mathbf{e}}_{xk}} = \frac{\mathbf{e}_{xk}^T\mathbf{W}\mathbf{V}\mathbf{W}\mathbf{e}_{xk}}{\mathbf{e}_{xk}^T\mathbf{W}\mathbf{e}_{xk}},$$

with $\mathbf{e}_{xk} = \mathbf{x}_k - \mathbf{X}_{(k)}\hat{\boldsymbol{\beta}}_{\text{SW}(k)}$ being the residual from SWLS regressing $\mathbf{x}_k$ on $\mathbf{X}_{(k)}$ and $\tilde{\mathbf{e}}_{xk} = \tilde{\mathbf{x}}_k - \tilde{\mathbf{X}}_{(k)}\hat{\boldsymbol{\beta}}_{\text{SW}(k)} = \mathbf{W}^{1/2}\mathbf{e}_{xk}$,

$$\rho_k = \frac{\tilde{\mathbf{x}}_k^T\tilde{\mathbf{x}}_k}{\tilde{\mathbf{x}}_k^T\tilde{\mathbf{V}}\tilde{\mathbf{x}}_k} = \frac{\mathbf{x}_k^T\mathbf{W}\mathbf{x}_k}{\mathbf{x}_k^T\mathbf{W}\mathbf{V}\mathbf{W}\mathbf{x}_k}$$

and $R^2_{\text{SW}(k)}$, defined in Appendix A, is the square of the multiple correlation from the weighted regression of the k^{th} column of $\mathbf{X}$ on the other columns. Hence, ζ_k and ρ_k depend on $\mathbf{W}$ and $\mathbf{V}$. The variance under orthogonality in (3) is inflated

$$\text{VIF}_k = \frac{\zeta_k\rho_k}{1 - R^2_{\text{SW}(k)}} \quad (6)$$

times when incorporating the other $p-1$ explanatory variables in SWLS. The model-based VIF in SWLS includes not only the multiple correlation coefficient $R^2_{\text{SW}(k)}$ but also two adjustment coefficients, ζ_k and ρ_k, that are not present in the OLS and WLS cases.

Using the singular value decomposition of $\tilde{\mathbf{V}}$, we can bound the factor $\zeta_k\rho_k$, which is the adjustment to the VIF in WLS. Based on the extrema of the ratio of quadratic forms (Lin 1984), the term ζ_k is bounded in the range of $\mu_{\min}(\tilde{\mathbf{V}}) \le \zeta_k \le \mu_{\max}(\tilde{\mathbf{V}})$, and ρ_k is bounded in the range of

$$\frac{1}{\mu_{\max}(\tilde{\mathbf{V}})} \le \rho_k \le \frac{1}{\mu_{\min}(\tilde{\mathbf{V}})},$$

where $\mu_{\min}(\tilde{\mathbf{V}})$ and $\mu_{\max}(\tilde{\mathbf{V}})$ are the minimum and maximum singular values of the matrix $\tilde{\mathbf{V}}$. Combining these results, the joint coefficient $\zeta_k\rho_k$ is bounded in the range of:

$$\frac{\mu_{\min}(\tilde{\mathbf{V}})}{\mu_{\max}(\tilde{\mathbf{V}})} \le \zeta_k\rho_k \le \frac{\mu_{\max}(\tilde{\mathbf{V}})}{\mu_{\min}(\tilde{\mathbf{V}})}.$$

Notice that when $\tilde{\mathbf{V}} = \mathbf{I}$, $\zeta_k = \rho_k = 1$ and (6) reduces to

$$\frac{1}{1 - R^2_{\text{SW}(k)}}\frac{\sigma^2}{\mathbf{x}_k^T\mathbf{W}\mathbf{x}_k},$$

which is the model variance of the WLS estimates when $\mathbf{V}$ is diagonal and $\mathbf{W}$ is correctly specified as $\mathbf{W} = \mathbf{V}^{-1}$. In that unusual case, the VIF currently computed by software packages will be appropriate for SWLS. However, rarely will it be reasonable to think that $\mathbf{W} = \mathbf{V}^{-1}$ in survey estimation. If $\tilde{\mathbf{V}} \ne \mathbf{I}$, then ζ_k and ρ_k are not equal to 1 and a specialized calculation of the VIF is still needed. When $\mathbf{V} = \mathbf{I}$, which is the usual application considered by analysts,

$$\tilde{\mathbf{V}} = \mathbf{W},\ \zeta_k = \frac{\tilde{\mathbf{e}}_{xk}^T\mathbf{W}\tilde{\mathbf{e}}_{xk}}{\tilde{\mathbf{e}}_{xk}^T\tilde{\mathbf{e}}_{xk}},\ \rho_k = \frac{\tilde{\mathbf{x}}_k^T\tilde{\mathbf{x}}_k}{\tilde{\mathbf{x}}_k^T\mathbf{W}\tilde{\mathbf{x}}_k}$$

and

$$\frac{\mu_{\min}(\tilde{\mathbf{V}})}{\mu_{\max}(\tilde{\mathbf{V}})} = \frac{w_{\min}}{w_{\max}},$$

where $w_{\min}$ is the minimum value of the survey weights and $w_{\max}$ is their maximum value. In this case, the range of $\zeta_k\rho_k$ is bounded by

$$\left[\frac{w_{\min}}{w_{\max}}, \frac{w_{\max}}{w_{\min}}\right].$$

When all the survey weights are constant, $\zeta_k\rho_k = 1$ and the VIF produced by standard software, $(1 - R^2_{\text{SW}})^{-1}$, does not need to be adjusted in SWLS; however, when the range of the survey weights is large, $\zeta_k\rho_k$ can be very small or large and can be either above or below 1. In this case the VIF produced by standard software is not appropriate and a special calculation is needed. These facts will be illustrated in our experimental studies.

The VIF in (6) is appropriate regardless of whether the model contains an intercept or not. An alternative version can also be written that assumes that an intercept is in the model when $\tilde{\mathbf{x}}_k$ is regressed on the other $\mathbf{x}$'s. The derivation of this form is in Liao (2010). We summarize the result below.

The variance of $\hat{\beta}_{\text{SW}_k}$ in a model M2 that includes an intercept and in which $\mathbf{x}_k$ is orthogonal to the other $\mathbf{x}$'s is:

$$\text{Var}_{M2}(\hat{\beta}_{\text{SW}_k}) = \frac{(\tilde{\mathbf{x}}_k - \tilde{\mathbf{1}}\bar{\tilde{x}}_k)^T \tilde{\mathbf{V}}(\tilde{\mathbf{x}}_k - \tilde{\mathbf{1}}\bar{\tilde{x}}_k)}{\text{SST}^2_{\text{SW}_{m(k)}}} \qquad (7)$$

where $\tilde{\mathbf{1}} = (w_1^{1/2}, ..., w_n^{1/2})$, $\bar{\tilde{x}}_k = \sum_{i\in s} w_i\, x_{ki} / \hat{N}$, $\hat{N} = \sum_{i\in s} w_i$, and $\text{SST}_{\text{SW}_{m(k)}} = \tilde{\mathbf{x}}_k^T \tilde{\mathbf{x}}_k - \hat{N}\bar{\tilde{x}}_k^2$. The variance of $\hat{\beta}_{\text{SW}_k}$ can then be rewritten as

$$\text{Var}_M(\hat{\beta}_{\text{SW}_k}) = \frac{\zeta_k\rho_{mk}}{1 - R^2_{\text{SW}m(k)}} \text{Var}_{M2}(\hat{\beta}_{\text{SW}_k}) \qquad (8)$$

where $R^2_{\text{SW}m(k)}$ is the SWLS R-square from regressing $\tilde{\mathbf{x}}_k$ on the $\mathbf{x}$'s in the remainder of $\tilde{\mathbf{X}}$ (excluding a column for the intercept). The term ζ_k was defined following (5) and

$$\rho_{mk} = \frac{(\tilde{\mathbf{x}}_k^T \tilde{\mathbf{x}}_k - \hat{N}\bar{\tilde{x}}_k^2)}{(\tilde{\mathbf{x}}_k - \tilde{\mathbf{1}}\bar{\tilde{x}}_k)^T \tilde{\mathbf{V}}(\tilde{\mathbf{x}}_k - \tilde{\mathbf{1}}\bar{\tilde{x}}_k)}.$$

Most software packages will consistently provide $(1 - R^2_{\text{SW}m(k)})^{-1}$ as the VIF as part of WLS regression output. Note that this is different from the VIF, $(1 - R^2_{\text{SW}(k)})^{-1}$, introduced in section 3.3 which does not assume that an intercept is retained in the model. Software packages generally do not supply $(1 - R^2_{\text{SW}(k)})^{-1}$.

Using arguments similar to those in the previous section, we can bound $\zeta_k\rho_{mk}$ by

$$\frac{\mu_{\min}(\tilde{\mathbf{V}})}{\mu_{\max}(\tilde{\mathbf{V}})} \le \zeta_k\rho_{mk} \le \frac{\mu_{\max}(\tilde{\mathbf{V}})}{\mu_{\min}(\tilde{\mathbf{V}})}.$$

The model variance of $\hat{\beta}_{\text{SW}_k}$ is inflated by

$$\text{VIF}_{mk} = \frac{\zeta_k\rho_{mk}}{1 - R^2_{\text{SW}m(k)}}$$

compared to its variance in the model (M2) with only the explanatory variable $\tilde{\mathbf{x}}_k$ and intercept. The new intercept-adjusted VIF_{mk} retains some properties of VIF_k in (6). When $\tilde{\mathbf{V}} = \mathbf{I}$, we have $\zeta_k = 1$, $\rho_{mk} = 1$ and the intercept-adjusted VIF in (8) for SWLS is equal to the conventional intercept-adjusted VIF: $(1 - R^2_{m(k)})^{-1}$. When $\mathbf{V} = \mathbf{I}$, we have $\tilde{\mathbf{V}} = \mathbf{W}$,

$$\zeta_k = \frac{\tilde{\mathbf{e}}_{xk}\mathbf{W}\tilde{\mathbf{e}}_{xk}}{\tilde{\mathbf{e}}_{xk}^T\tilde{\mathbf{e}}_{xk}},\ \rho_{mk} = \frac{(\tilde{\mathbf{x}}_k^T\tilde{\mathbf{x}}_k - \hat{N}\bar{\tilde{x}}_k^2)}{(\tilde{\mathbf{x}}_k - \tilde{\mathbf{1}}\bar{\tilde{x}}_k)^T\mathbf{W}(\tilde{\mathbf{x}}_k - \tilde{\mathbf{1}}\bar{\tilde{x}}_k)}$$

and

$$\frac{\mu_{\min}(\tilde{\mathbf{V}})}{\mu_{\max}(\tilde{\mathbf{V}})} = \frac{w_{\min}}{w_{\max}}.$$

The range of $\zeta_k\rho_{mk}$ also depends on the range of survey weights as did $\zeta_k\rho_k$.

3.4 Estimating the VIF for a model with stratified clustering when $\mathbf{V}$ is unknown

In the previous sections, we used model-based arguments to derive VIFs. The VIFs contain terms, $\tilde{\mathbf{V}}$ in particular, that are unknown and must be estimated. In this section, we construct estimators of the components of the VIFs, again using model-based arguments. However, a standard, design-based linearization variance estimator also estimates the model variance, as shown below, and supplies the components needed to estimate the VIF. In the remainder of this section, we will present estimators that are appropriate for a model that has a stratified clustered covariance structure.

Suppose that in a stratified multistage sampling design, there are $h = 1, ..., H$ strata in the population, $i = 1, ..., N_h$ clusters in the corresponding stratum h and $t = 1, ..., M_{hi}$ units in cluster hi. We select $i = 1, ..., n_h$ clusters in stratum h and $t = 1, ..., m_{hi}$ units in cluster hi. Denote the set of sample clusters in stratum h by s_h and the sample of units in cluster hi as s_{hi}. The total number of sample units in stratum h is $m_h = \sum_{i\in s_h} m_{hi}$, and the total in the sample is $m = \sum_{h=1}^H m_h$. Clusters are assumed to be selected with replacement within strata and independently between strata. Consider this model:

$$\begin{aligned} &E_M(Y_{hit}) = \mathbf{x}_{hit}^T\boldsymbol{\beta} \\ &\qquad h = 1, \ldots, H, \quad i = 1, \ldots, N_h, \quad t = 1, \ldots, M_{hi} \\ &\text{Cov}_M(Y_{hit}, Y_{h'i't'}) = 0 \\ &\qquad h \ne h', \text{ or, } h = h' \text{ and } i \ne i'. \end{aligned} \qquad (9)$$

Units within each cluster are assumed to be correlated but the particular correlation of the covariances does not have to be specified for this analysis. The estimator of the regression parameter is:

$$\hat{\boldsymbol{\beta}}_{\text{SW}} = \sum_{h=1}^H \sum_{i\in s_h} \mathbf{A}^{-1}\mathbf{X}_{hi}^T\mathbf{W}_{hi}\mathbf{Y}_{hi} \qquad (10)$$

where $\mathbf{X}_{hi}$ is the $m_{hi} \times p$ matrix of covariates for sample units in cluster hi, $\mathbf{W}_{hi} = \text{diag}(w_t)$, $t \in s_{hi}$ is the diagonal

matrix of survey weights for cluster hi and $\mathbf{Y}_{hi}$ is the $m_{hi} \times 1$ vector of response variables in cluster hi. The model variance of $\hat{\boldsymbol{\beta}}_{\mathrm{SW}}$ is:

$$\mathrm{Var}_M(\hat{\boldsymbol{\beta}}_{\mathrm{SW}}) = \mathbf{A}^{-1}\left[\sum_{h=1}^{H}\sum_{i\in s_h}\mathbf{X}_{hi}^T\mathbf{W}_{hi}\mathbf{V}_{hi}\mathbf{W}_{hi}\mathbf{X}_{hi}\right]\mathbf{A}^{-1} = \mathbf{A}^{-1}\left[\sum_{h=1}^{H}\mathbf{X}_h^T\mathbf{W}_h\mathbf{V}_h\mathbf{W}_h\mathbf{X}_h\right]\mathbf{A}^{-1}, \quad (11)$$

where $\mathbf{V}_{hi} = \mathrm{Var}_M(\mathbf{Y}_{hi})$ and $\mathbf{V}_h = \mathrm{Blkdiag}(\mathbf{V}_{hi})$, $i \in s_h$. Expression (11) is a special case of (2) with $\mathbf{X}^T = (\mathbf{X}_1^T, \mathbf{X}_2^T, ..., \mathbf{X}_H^T)$, $\mathbf{X}_h$ is the $m_h \times p$ matrix of covariates for sample units in stratum h, $\mathbf{W} = \mathrm{diag}(\mathbf{W}_{hi})$, for $h = 1, ..., H$ and $i \in s_h$ and $\mathbf{V} = \mathrm{Blkdiag}(\mathbf{V}_h)$.

Denote the cluster-level residuals as a vector, $\mathbf{e}_{hi} = \mathbf{Y}_{hi} - \mathbf{X}_{hi}\hat{\boldsymbol{\beta}}_{\mathrm{SW}}$. A design-based linearization estimator is:

$$\mathrm{var}_L(\hat{\boldsymbol{\beta}}_{\mathrm{SW}}) = \mathbf{A}^{-1}\left[\sum_{h=1}^{H}\frac{n_h}{n_h-1}\sum_{i\in s_h}(\mathbf{z}_{hi}-\bar{\mathbf{z}}_h)(\mathbf{z}_{hi}-\bar{\mathbf{z}}_h)^T\right]\mathbf{A}^{-1} = \mathbf{A}^{-1}\left[\sum_{h=1}^{H}\frac{n_h}{n_h-1}\left(\sum_{i\in s_h}\mathbf{z}_{hi}\mathbf{z}_{hi}^T - n_h\bar{\mathbf{z}}_h\bar{\mathbf{z}}_h^T\right)\right]\mathbf{A}^{-1}, \quad (12)$$

where

$$\bar{\mathbf{z}}_h = \frac{1}{n_h}\sum\nolimits_{i\in s_h}\mathbf{z}_{hi}$$

and $\mathbf{z}_{hi} = \mathbf{X}_{hi}^T\mathbf{W}_{hi}\mathbf{e}_{hi}$ with $\mathbf{e}_{hi} = \mathbf{Y}_{hi} - \mathbf{X}_{hi}\hat{\boldsymbol{\beta}}_{\mathrm{SW}}$. This expression can be reduced to the formula for a single-stage stratified design when the cluster sample sizes are all equal to 1, $m_{hi} = 1$. Expression (12) is used by the Stata and SUDAAN packages, among others. The estimator $\mathrm{var}_L(\hat{\boldsymbol{\beta}}_{\mathrm{SW}})$ is consistent and approximately design-unbiased under a design where clusters are selected with replacement (Fuller 2002). Li (2007a, b) showed that (12) is also an approximately model-unbiased estimator under model (11).

The term in brackets in (12) serves as an estimator of the matrix $\mathbf{B}$ in expression (2). The components of $\mathrm{var}_L(\hat{\boldsymbol{\beta}}_{\mathrm{SW}})$ can be used to construct estimators of ζ_k and ρ_k in (5) and ρ_{mk} in (8). In particular,

$$\hat{\zeta}_k = \frac{\mathbf{e}_{xk}^T\mathbf{W}\hat{\mathbf{V}}\mathbf{W}\mathbf{e}_{xk}}{\mathbf{e}_{xk}^T\mathbf{W}\mathbf{e}_{xk}}, \quad (13)$$

where

$$\hat{\mathbf{V}} = \mathrm{Blkdiag}\left[\frac{n_h}{n_h-1}\left(\hat{\mathbf{V}}_h - \frac{1}{n_h}\mathbf{e}_h\mathbf{e}_h^T\right)\right],\ h = 1, 2, ..., H,$$

with $\hat{\mathbf{V}}_h = \mathrm{Blkdiag}(\mathbf{e}_{hi}\mathbf{e}_{hi}^T)$ and

$$\hat{\rho}_k = \frac{\mathbf{x}_k^T\mathbf{W}\mathbf{x}_k}{\mathbf{x}_k^T\mathbf{W}\hat{\mathbf{V}}\mathbf{W}\mathbf{x}_k},$$

with $\mathbf{e}_{xk} = \mathbf{x}_k - \mathbf{X}_{(k)}\hat{\boldsymbol{\beta}}_{\mathrm{SW}(k)}$. The estimate of $\hat{\rho}_{mk}$, defined following (8), is

$$\hat{\rho}_{mk} = \frac{(\mathbf{x}_k^T\mathbf{W}\mathbf{x}_k - \hat{N}\,\bar{\tilde{x}}_k^2)}{(\mathbf{x}_k - \mathbf{1}\bar{\tilde{x}}_k)^T\mathbf{W}\hat{\mathbf{V}}\mathbf{W}(\mathbf{x}_k - \mathbf{1}\bar{\tilde{x}}_k)}. \quad (14)$$

Given these component estimators VIF_k is estimated by

$$\widehat{\mathrm{VIF}}_k = \frac{\hat{\zeta}_k\hat{\rho}_k}{1 - R^2_{\mathrm{SW}(k)}}$$

and VIF_{mk} is estimated by

$$\widehat{\mathrm{VIF}}_{mk} = \frac{\hat{\zeta}_k\,\hat{\rho}_{mk}}{1 - R^2_{\mathrm{SW}m(k)}}.$$

4. Experimental study

We will now illustrate the proposed, modified collinearity diagnostics and investigate their behavior using dietary intake data from the National Health and Nutrition Examination Survey (NHANES) 2007-2008 (http://www.cdc.gov/nchs/nhanes/nhanes2007-2008/datadoc_changes_0708.htm). The dietary intake data are used to estimate the types and amounts of foods and beverages consumed during the 24-hour period prior to the interview (midnight to midnight), and to estimate intakes of energy, nutrients, and other food components from those foods and beverages. NHANES uses a complex, multistage, probability sampling design. Oversampling of certain population subgroups is done to increase the reliability and precision of health status indicator estimates for these groups. Among the respondents who received the in-person interview in the mobile examination center (MEC), around 94% provided complete dietary intakes. The survey weights in this data were constructed by taking MEC sample weights and further adjusting for the additional nonresponse and the differential allocation by day of the week for the dietary intake data collection. These weights are more variable than the MEC weights. The data set used in our study is a subset of 2007-2008 data composed of female respondents aged 26 to 40. Observations with missing values in the selected variables are excluded from the sample which finally contains 672 complete respondents. The final weights in our sample range from 6,028 to 330,067, with a ratio of 55:1. The U.S. National Center for Health Statistics recommends that the design of the sample is approximated by the stratified selection with replacement of 32 PSUs from 16 strata, with 2 PSUs within each stratum.

For this empirical study, a linear regression of body weight(kg) is fitted using survey weighted least squares. The predictor variables considered include age, Black(race) and

nine daily total nutrition intake variables, which are calorie(100kcal), protein(100gm), carbohydrate(100gm), sugar(100gm), total fat(100gm), total saturated fatty acids(100gm), total monounsaturated fatty acids(100gm), total polyunsaturated fatty acids(100gm) and alcohol(100gm). All the daily total nutrition intake variables are correlated with each other to different degrees as shown in Figure 1.

Three regression methods compared in this study. The first one uses *ordinary least squares* (OLS) method and ignores sampling complexities including the weighting. The second one uses *weighted least squares* (WLS), which incorporates the survey weights by assuming $\mathbf{V} = \mathbf{W}^{-1}$ but ignores all sampling complexities. The third one is *survey weighted least squares* (SWLS), which uses the actual complex sampling design as described in section 3.4. The weight matrices, coefficient variance estimators and collinearity diagnostics of these three methods are listed in Table 1.

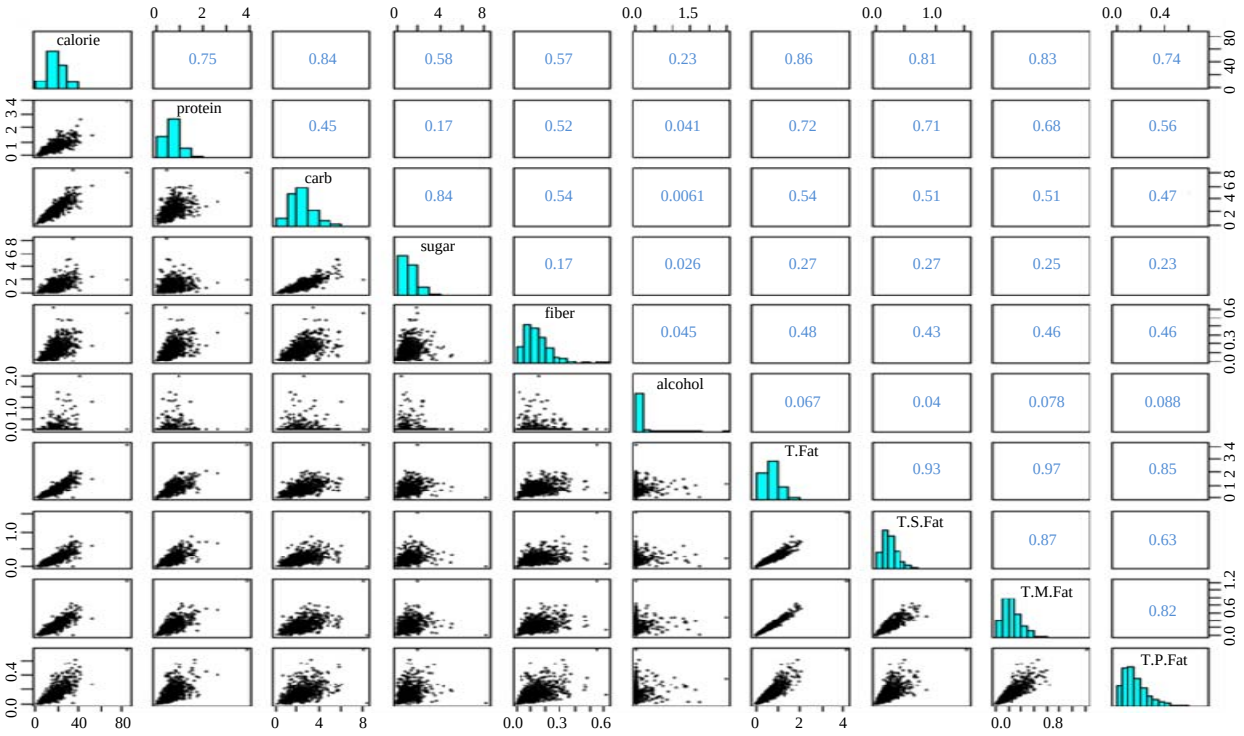
The results from fitting the model using three different regression methods are displayed in Table 2. The model with all the predictors is shown in the upper part of the table. In the lower tier of the table, a reduced model with less of the near-dependency problem is fitted with only three predictors: age, Black and calorie. In the reduced model, the value of the coefficient for calorie is positive and significant when WLS or SWLS is used, which seems logical and reflects the anticipated positive relationship between a respondent's body weight and her daily total calorie intake. However, when the other total nutrition intake variables are included in the model, the value of the calorie coefficient is negative and not significant due to its inflated variance. This is a typical example in which the variance of a coefficient is inflated, and its sign is illogical due to collinearity.

Table 3 reports the VIF values when the three different regression methods are used. The VIF formulas for these regression methods are listed in Table 1. When all the predictors are included in the model, calorie has the largest VIF values in all the regressions due to its high near-dependency with all the other total nutrition intake variables. As shown in Table 1, the VIF in SWLS can be obtained by multiplying the VIF from WLS with the adjustment coefficient $\zeta_k\rho_k$. In Table 3, the adjustment coefficients $\zeta_k\rho_k$ for all the non-fat total nutrition intake variables are all less than 1, especially the one for carbohydrate which is 0.46. This indicates that the VIF values for these variables in SWLS are much smaller than the ones in WLS and the collinearity among predictors in the model has less impact on the coefficient estimation when using SWLS, compared to using WLS. But for the fat-related nutrition intake variables, their $\zeta_k\rho_k$ are all larger than 1. Thus, the collinearity among the fat-related nutrition intake variables is more harmful to the coefficient estimation in SWLS than in WLS. To take a closer look at this problem, we also fitted a model that only contains two nutrition intake variables: total fat and total monounsatruated fatty acids. The SWLS VIF values are three times as large as the ones from OLS or WLS for these two nutrition variables. If an analyst is analyzing this survey data using SWLS but uses the unadjusted VIF values provided by standard statistical packages for either OLS (as shown in the first column) or WLS (as shown in the second column), the unadjusted VIFs will give somewhat misleading judgements on the severity of collinearity in this model. In summary, although the estimated slopes and predictions in regression using WLS and SWLS are the same, the VIFs can be underestimated or overestimated if survey complexities are ignored.

Table 1
Regression methods and their collinearity diagnostic statistics used in this experimental study

Regression Type	Weight Matrix $\mathbf{W}^a$	Variance Estimation of $\hat{\boldsymbol{\beta}}$	VIF fomula
OLS	$\mathbf{I}$	$\hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}$	$\text{VIF} = \dfrac{1}{1-R^2_{m(k)}}$ [b]
WLS	$\mathbf{W}^c$	$\hat{\sigma}^2(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}$	$\text{VIF} = \dfrac{1}{1-R^2_{\text{SW}m(k)}}$
SWLS	$\mathbf{W}$	$\hat{\sigma}^2(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\hat{\mathbf{V}}\mathbf{W}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}$	$\text{VIF} = \dfrac{\hat{\zeta}_k\hat{\rho}_{mk}}{1-R^2_{\text{SW}m(k)}}$
		with	with $\hat{\zeta}_k = \dfrac{\mathbf{e}_{xk}^T\mathbf{W}\hat{\mathbf{V}}\mathbf{W}\mathbf{e}_{xk}}{\mathbf{e}_{xk}^T\mathbf{W}\mathbf{e}_{xk}}$,
		$\hat{\mathbf{V}} = \sum_{h=1}^{H}\dfrac{n_h}{n_h-1}\left[\text{Blkdiag}(\mathbf{e}_{hi}\mathbf{e}_{hi}^T) - \dfrac{1}{n_h}\mathbf{e}_h\mathbf{e}_h^T\right]$	$\hat{\rho}_{mk} = \dfrac{(\tilde{\mathbf{x}}_k^T\tilde{\mathbf{x}}_k - \hat{N}\bar{\tilde{x}}_k^2)}{(\tilde{\mathbf{x}}_k - \tilde{\mathbf{1}}\bar{\tilde{x}}_k)^T\hat{\mathbf{V}}(\tilde{\mathbf{x}}_k - \tilde{\mathbf{1}}\bar{\tilde{x}}_k)}$

[a] In all the regression models, the parameters are estimated by: $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y}$.
[b] $R^2_{m(k)}$ is the OLS R-square from regressing $\mathbf{x}_k$ on the $\mathbf{x}$'s in the remainder of $\mathbf{X}$ (excluding a column for the intercept).
[c] $\mathbf{W}$ is the diagonal matrix with survey weights w_i on the main diagonal.



[a] T.Fat: total fat;
T.S.Fat: total saturated fatty acid;
T.M.Fat: total monounsaturated fatty acid;
T.P.Fat: total polyunsaturated fatty acid.

Figure 1 Pairwise scatterplots and correlation coefficients of nutrition variables[a]

Table 2
Parameter estimates with their associated standard errors using three different regression methods

	Full Model					
	OLS		WLS		SWLS	
Variable	**Beta**	**SE.**	**Beta**	**SE.**	**Beta**	**SE.**
Intercept	63.90***[a]	6.95	67.47***	6.36	67.47***	8.76
Age	0.26	0.19	0.08	0.18	0.08	0.25
Black	10.39***	2.07	10.59***	2.38	10.59***	2.20
Calorie	-6.41	5.76	-8.19	5.56	-8.19	5.75
Protein	25.72	24.76	40.98	23.60	40.98	25.38
Carbohydrate	26.67	23.93	32.31	22.96	32.31	22.65
Sugar	-1.90	3.06	-0.30	2.82	-0.30	4.06
Fiber	-41.17	20.23	-34.20	17.98	-34.20	19.05
Alcohol	38.84	39.45	49.37	38.28	49.37	40.10
Total Fat	150.25*	69.53	161.78*	72.12	161.78	94.76
Total Saturated Fatty Acids	-113.20*	49.81	-101.40	56.26	-101.40	82.71
Total Monounsaturated Fatty Acids	-72.05	48.03	-92.44	51.52	-92.44	83.51
Total Polyunsaturated Fatty Acids	-92.60*	46.13	-75.55	51.16	-75.55	78.76
	Reduced Model					
	OLS		WLS		SWLS	
Variable	**Beta**	**SE.**	**Beta**	**SE.**	**Beta**	**SE.**
Intercept	62.26***	6.88	67.52***	6.29	67.52***	8.48
Age	0.27	0.19	0.07	0.18	0.07	0.25
Black	12.54***	1.98	11.74***	2.32	11.74***	2.05
Calorie	0.15	0.10	0.23*	0.09	0.23*	0.10

[a] p values of significance: * $p = 0.05$; ** $p = 0.01$; *** $p = 0.005$.

Table 3
VIF values using three different regression methods

	Full Model			
	OLS	WLS	SWLS	
Variable	**VIF**	**VIF**	**VIF**	$\zeta_k \rho_k$
Age	1.02	1.03	0.96	0.94
Black	1.10	1.07	1.12	1.05
Calorie	3,411.61	3,562.70	2,740.83	0.77
Protein	123.12	127.35	103.50	0.81
Carbohydrate	1,074.87	1,007.40	462.08	0.46
Sugar	8.37	7.03	4.87	0.69
Fiber	4.59	3.94	2.37	0.60
Alcohol	120.56	115.67	89.92	0.78
Total Fat	1,190.24	1,475.27	2,513.69	1.70
Total Saturated Fatty Acids	76.80	112.61	202.91	1.80
Total Monounsaturated Fatty Acids	82.37	107.34	286.24	2.67
Total Polyunsaturated Fatty Acids	34.73	49.45	118.21	2.39
	Reduced Model			
	OLS	WLS	SWLS	
Variable	**VIF**	**VIF**	**VIF**	$\zeta_k \rho_k$
Age	1.00	1.00	0.98	0.98
Black	1.02	1.01	0.97	0.96
Total Fat	20.10	20.22	63.15	3.12
Total Monounsaturated Fatty Acids	20.16	20.26	61.57	3.04
	Reduced Model			
	OLS	WLS	SWLS	
Variable	**VIF**	**VIF**	**VIF**	$\zeta_k \rho_k$
Age	1.00	1.00	0.98	0.97
Black	1.00	1.03	1.00	1.00
Calorie	1.00	1.01	0.96	0.95

5. Conclusion

Regression diagnostics need to be adapted to be appropriate for models estimated from survey data to account for the use of weights and design features like stratification and clustering. In this paper we developed a new formulation for a variance inflation factor (VIF) appropriate for linear models. A VIF measures the amount by which the variance of a parameter estimator is inflated due to predictor variables being correlated with each other, rather than being orthogonal. Although survey-weighted regression slope estimates can be obtained from weighted least squares procedures in standard software packages, the VIFs produced by the non-survey routines are incorrect. The complex sample VIF is equal to the VIF from weighted least squares times an adjustment factor. The adjustment factor is positive but can be either larger or smaller than 1, depending on the nature of the data being analyzed.

In an empirical study, we illustrated the application of our new approach using data from the 2007-2008 National Health and Nutrition Examination Survey. We provided a simple example of how the collinearity among predictors affects the estimation of coefficients in linear regression and demonstrated that although the estimated coefficients (and fitted values) are the same when weighted least squares or survey-weighted least squares are used, their estimated variances and VIF values (reflecting the impact of collinearity on coefficient estimation) can be different.

The goals of an analysis must be considered in deciding how to use VIFs. If prediction is the main objective, then including collinear variables or selecting incorrect variables is less of a concern. If more substantive conclusions are desired, then an analyst should consider which variables should logically be included as predictors rather than relying on some automatic algorithm for variable selection. VIFs are a useful tool for identifying predictors whose estimated coefficients have variances that are unnecessarily large. Although VIFs might be considered as a tool for automatic variable selection, simulations in Liao (2010), not reported here, show that using VIFs is not a reliable way of identifying a true underlying model.

Acknowledgements

The authors thank the associate editor and referees whose comments led to important improvements.

Appendix A

Derivation of g^{kk}

Similar to the derivation of conventional OLS VIF in Theil (1971), the sum of squares and cross products matrix $\mathbf{A} = \tilde{\mathbf{X}}^T\tilde{\mathbf{X}}$, which can be partitioned as

$$\mathbf{A}_{p\times p} = \begin{pmatrix} \tilde{\mathbf{x}}_k^T\tilde{\mathbf{x}}_k & \tilde{\mathbf{x}}_k^T\tilde{\mathbf{X}}_{(k)} \\ \tilde{\mathbf{X}}_{(k)}^T\tilde{\mathbf{x}}_k & \tilde{\mathbf{X}}_{(k)}^T\tilde{\mathbf{X}}_{(k)} \end{pmatrix} \quad (15)$$

where the columns of $\tilde{\mathbf{X}}$ are reordered so that $\tilde{\mathbf{X}} = (\tilde{\mathbf{x}}_k\ \tilde{\mathbf{X}}_{(k)})$ with $\tilde{\mathbf{X}}_{(k)}$ being the $n\times(p-1)$ matrix containing all columns except the k^{th} column of $\tilde{\mathbf{X}}$.

Using the formula for the inverse of a partitioned matrix, the upper-left element of $\mathbf{A}^{-1}$ can be expressed as:

$$\begin{aligned} a^{kk} = \mathbf{i}_k^T\mathbf{A}^{-1}\mathbf{i}_k &= \mathbf{i}_k^T(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\mathbf{i}_k \\ &= \frac{1}{\left(1-R^2_{\text{SW}_{(k)}}\right)\text{SST}_{\text{SW}_{(k)}}} \\ &= \frac{1}{\left(1-R^2_{\text{SW}_{(k)}}\right)\tilde{\mathbf{x}}_k^T\tilde{\mathbf{x}}_k} \end{aligned} \quad (16)$$

where

$$R^2_{\text{SW}_{(k)}} = \frac{\hat{\beta}^T_{\text{SW}_{(k)}}\tilde{\mathbf{X}}^T_{(k)}\tilde{\mathbf{X}}_{(k)}\hat{\beta}_{\text{SW}_{(k)}}}{\text{SST}_{\text{SW}_{(k)}}},$$

with $\hat{\beta}_{\text{SW}_{(k)}} = (\tilde{\mathbf{X}}^T_{(k)}\tilde{\mathbf{X}}_{(k)})^{-1}\tilde{\mathbf{X}}^T_{(k)}\tilde{\mathbf{x}}_k$, is the coefficient of determination corresponding to the regression of $\mathbf{x}_k$ on the $p-1$ other explanatory variables. The term $\text{SST}_{\text{SW}_{(k)}} = \tilde{\mathbf{x}}_k^T\tilde{\mathbf{x}}_k$, is the total sum of squares in this regression.

The term $(1-R^2_{\text{SW}_{(k)}})^{-1}$ in (16) is the VIF that will be produced by standard statistical packages when a weighted least squares regression is run. Under the model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim (0, \sigma^2\mathbf{W}^{-1})$, expression (16) is equal to $\text{Var}_M(\hat{\beta}_{\text{SW}_k})/\sigma^2$. However, this is not appropriate for survey-weighted least squares regressions because the variance of $\hat{\boldsymbol{\beta}}_{\text{SW}}$ has the more complex form in (2).

The matrix $\mathbf{G} = \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$ can be expressed as:

$$\mathbf{G} = \begin{pmatrix} a^{kk} & \mathbf{a}^{k(k)} \\ \mathbf{a}^{(k)k} & \mathbf{A}^{(k)(k)} \end{pmatrix}\begin{pmatrix} b_{kk} & \mathbf{b}_{k(k)} \\ \mathbf{b}_{(k)k} & \mathbf{B}_{(k)(k)} \end{pmatrix}\begin{pmatrix} a^{kk} & \mathbf{a}^{k(k)} \\ \mathbf{a}^{(k)k} & \mathbf{A}^{(k)(k)} \end{pmatrix} \quad (17)$$

where the inverse matrix is $\mathbf{A}^{-1} = [a^{hk}]$, $h, k = 1, ..., p$, $\mathbf{a}^{k(k)}$ is defined as the k^{th} row of $\mathbf{A}^{-1}$ excluding a^{kk}, $(a^{k1}, ..., a^{k(k-1)}, a^{k(k+1)}, ..., a^{kp})$, $\mathbf{a}^{(k)k} = [\mathbf{a}^{k(k)}]^T$ and $\mathbf{A}^{(k)(k)}$ is defined as the $(k-1)\times(k-1)$ part of matrix $\mathbf{A}^{-1}$ excluding the k^{th} row and column. The partitioned version of $\mathbf{B}$ is

$$\mathbf{B} = \begin{pmatrix} b_{kk} & \mathbf{b}_{k(k)} \\ \mathbf{b}_{(k)k} & \mathbf{B}_{(k)(k)} \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{x}}_k^T\tilde{\mathbf{V}}\tilde{\mathbf{x}}_k & \tilde{\mathbf{x}}_k^T\tilde{\mathbf{V}}\tilde{\mathbf{X}}_{(k)} \\ \tilde{\mathbf{X}}^T_{(k)}\tilde{\mathbf{V}}\tilde{\mathbf{x}}_k & \tilde{\mathbf{X}}^T_{(k)}\tilde{\mathbf{V}}\tilde{\mathbf{X}}_{(k)} \end{pmatrix}. \quad (18)$$

By virtue of the symmetry of $\mathbf{A}$ and $\mathbf{B}$, the k^{th} diagonal element of $\mathbf{G}$ is

$$g^{kk} = a^{kk}(a^{kk}b_{kk} + 2\mathbf{b}_{k(k)}\mathbf{a}^{(k)k}) + \mathbf{a}^{(k)kT}\mathbf{B}_{(k)(k)}\mathbf{a}^{(k)k}. \quad (19)$$

Using the partitioned inverse of matrix $\mathbf{A}$, which represents $(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}$, it can be shown that

$$\mathbf{a}^{(k)k} = -a^{kk}(\tilde{\mathbf{X}}^T_{(k)}\tilde{\mathbf{X}}_{(k)})^{-1}\tilde{\mathbf{X}}_{(k)}\tilde{\mathbf{x}}_k = -a^{kk}\hat{\boldsymbol{\beta}}_{\text{SW}(k)}. \quad (20)$$

Substituting $a^{(k)k}$ in (19), g^{kk} can be compactly expressed in terms of a^{kk}, $\hat{\boldsymbol{\beta}}_{\text{SW}(k)}$ and the lower right component of matrix $\mathbf{B}$:

$$\begin{aligned} g^{kk} &= (a^{kk})^2\left(b_{kk} - 2\mathbf{b}_{k(k)}\hat{\boldsymbol{\beta}}_{\text{SW}(k)} + \hat{\boldsymbol{\beta}}^T_{\text{SW}(k)}\mathbf{B}_{(k)(k)}\hat{\boldsymbol{\beta}}_{\text{SW}(k)}\right) \\ &= a^{kk}\times\frac{1}{1-R^2_{\text{SW}(k)}}\frac{1}{\tilde{\mathbf{x}}_k^T\tilde{\mathbf{x}}_k} \\ &\quad\times\left(\tilde{\mathbf{x}}_k^T\tilde{\mathbf{V}}\tilde{\mathbf{x}}_k - 2\tilde{\mathbf{x}}_k^T\tilde{\mathbf{V}}\tilde{\mathbf{X}}_{(k)}\hat{\boldsymbol{\beta}}_{\text{SW}(k)} + \hat{\boldsymbol{\beta}}^T_{\text{SW}(k)}\tilde{\mathbf{X}}_{(k)}\tilde{\mathbf{V}}\tilde{\mathbf{X}}_{(k)}\hat{\boldsymbol{\beta}}_{\text{SW}(k)}\right) \\ &= a^{kk}\times\frac{\left(\tilde{\mathbf{x}}_k - \tilde{\mathbf{X}}_{(k)}\hat{\boldsymbol{\beta}}_{\text{SW}(k)}\right)^T\tilde{\mathbf{V}}\left(\tilde{\mathbf{x}}_k - \tilde{\mathbf{X}}_{(k)}\hat{\boldsymbol{\beta}}_{\text{SW}(k)}\right)}{\left(1-R^2_{\text{SW}(k)}\right)\tilde{\mathbf{x}}_k^T\tilde{\mathbf{x}}_k} \\ &= a^{kk}\times\frac{\left(\tilde{\mathbf{x}}_k - \tilde{\mathbf{X}}_{(k)}\hat{\boldsymbol{\beta}}_{\text{SW}(k)}\right)^T\tilde{\mathbf{V}}\left(\tilde{\mathbf{x}}_k - \tilde{\mathbf{X}}_{(k)}\hat{\boldsymbol{\beta}}_{\text{SW}(k)}\right)}{\left(\tilde{\mathbf{x}}_k - \tilde{\mathbf{X}}_{(k)}\hat{\boldsymbol{\beta}}_{\text{SW}(k)}\right)^T\left(\tilde{\mathbf{x}}_k - \tilde{\mathbf{X}}_{(k)}\hat{\boldsymbol{\beta}}_{\text{SW}(k)}\right)} \\ &= \frac{1}{1-R^2_{\text{SW}(k)}}\frac{1}{\tilde{\mathbf{x}}_k^T\tilde{\mathbf{x}}_k}\frac{\tilde{\mathbf{e}}_{xk}^T\tilde{\mathbf{V}}\tilde{\mathbf{e}}_{xk}}{\tilde{\mathbf{e}}_{xk}^T\tilde{\mathbf{e}}_{xk}}, \end{aligned} \quad (21)$$

where $\tilde{\mathbf{e}}_{xk} = \tilde{\mathbf{x}}_k - \tilde{\mathbf{X}}_{(k)}\hat{\boldsymbol{\beta}}_{\text{SW}(k)}$ is the residual from regressing $\tilde{\mathbf{x}}_k$ on $\tilde{\mathbf{X}}_{(k)}$.

References

Belsley, D.A. (1984). Collinearity and forecasting. *Journal of Forecasting*, 38, 73-93.

Belsley, D.A. (1991). *Conditioning Diagnostics, Collinearity and Weak Data in Regression*. New York: John Wiley & Sons, Inc.

Belsley, D.A., Kuh, E. and Welsch, R.E. (1980). *Regression Diagnostics*: *Identifying Influential Data and Sources of Collinearity*. Wiley Series in Probability and Statistics. New York: Wiley Interscience.

Elliot, M. (2007). Bayesian weight trimming for generalized linear regression models. *Survey Methodology*, 33, 23-34.

Farrar, D.E., and Glauber, R.R. (1967). Multicollinearity in regression analysis. *Review of Economics and Statistics*, 49, 92-107.

Fox, J. (1984). *Linear Statistical Models and Related Methods, With Applications to Social Research*. New York: John Wiley & Sons, Inc.

Fox, J., and Monette, G. (1992). Generalized collinearity diagnostics. *Journal of the American Statistical Association*, 87(417), 178-183.

Fuller, W.A. (2002). Regression estimation for survey samples. *Survey Methodology*, 28(1), 5-23.

Kmenta, J. (1986). *Elements of Econometrics*. New York: Macmillan, 2nd Ed.

Li, J. (2007a). Linear regression diagnostics in cluster samples. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 3341-3348.

Li, J. (2007b). Regression diagnostics for complex survey data. Unpublished doctoral dissertation, University of Maryland. Available at http://drum.lib.umd.edu/bitstream/1903/7598/1/umi-umd-4863.pdf.

Li, J., and Valliant, R. (2009). Survey weighted hat matrix and leverages. *Survey Methodology*, 35(1), 15-24.

Li, J., and Valliant, R. (2011). Linear regression influence diagnostics for unclustered survey data. *Journal of Official Statistics*, 20, 99-119.

Liao, D. (2010). *Collinearity Diagnostics for Complex Survey Data*. Unpublished doctoral dissertation, University of Maryland. Available at http://drum.lib.umd.edu/bitstream/1903/10881/1/Liao_umd_0117E_11537.pdf.

Lin, C. (1984). Extrema of quadratic forms and statistical applications. *Communications in Statistics-Theory and Methods*, 13, 1517-1520.

Neter, J., Kutner, M., Wasserman, W. and Nachtsheim, C. (1996). *Applied Linear Statistical Models*. New York: McGraw-Hill/Irwin, 4th Ed.

Simon, S.D., and Lesage, J.P. (1988). The impact of collinearity involving the intercept term on the numerical accuracy of regression. *Computer Science in Economics and Management*, 1, 137-152.

Smith, G. (1974). Multicollinearity and forecasting. Yale Univeristy Cowles Foundation Disscussion Paper No.383. Available at http://cowles.econ.yale.edu/P/cd/d03b/d0383.pdf.

Steward, G.W. (1987). Collinearity and least squares regression. *Statistical Science*, 2(1), 68-84.

Theil, H. (1971). *Principles of Econometrics*. New York: John Wiley & Sons, Inc.